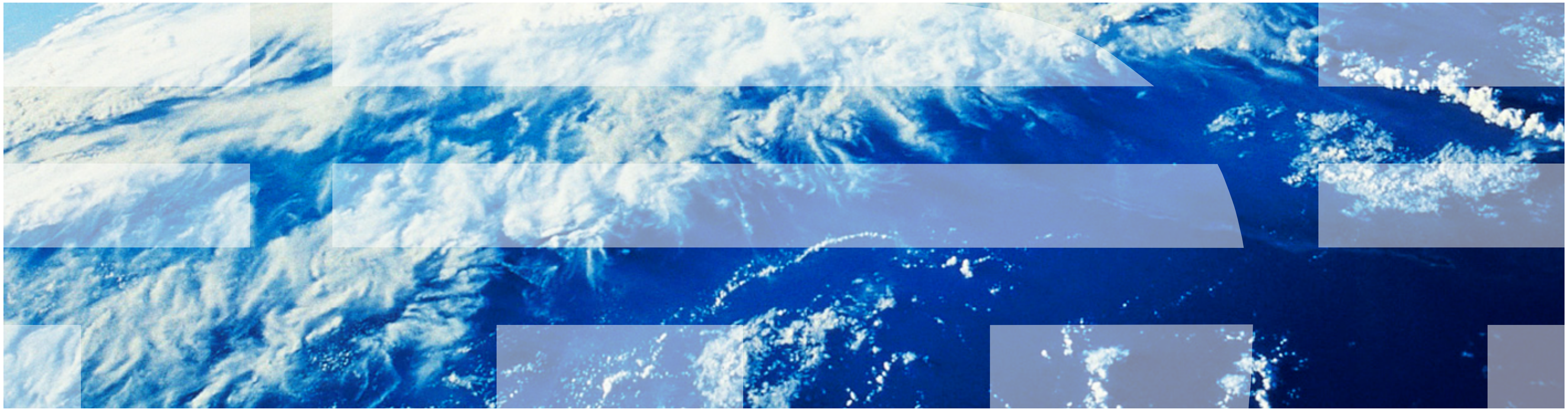
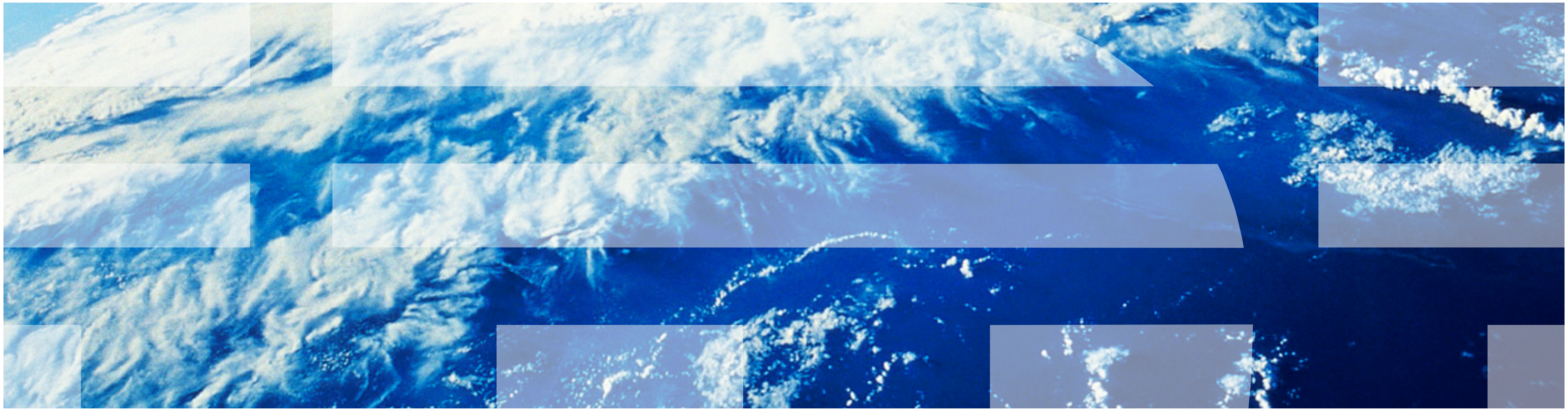
Computer Systems for Data Science
Topic 9

Data Security

Compliance



Security Goals



Four security goals

- Confidentiality
  - My secret data was not leaked/stolen
    - E.g., the government, my spouse, a hacker did not gain access to my private information
- Integrity
  - Data / system was not tampered with
    - E.g., nobody added another ‘0’ in their bank account entry
    - E.g., nobody changed the content of my folder without me noticing
- Availability
  - Data or system will be available when needed
    - E.g., my website/service is protected from a malicious hacker trying to take them with a denial of service attack
- Authenticity
  - Belief in the source of the data
    - E.g., this new COVID study came from a reputable journal
    - E.g., the person communicating with me is who they say they are

Cryptography can help address most of these goals

- **Confidentiality**
  - My secret data was not leaked/stolen
    - E.g., the government, my spouse, a hacker did not gain access to my private information
- **Integrity**
  - Data / system was not tampered with
    - E.g., nobody added another ‘0’ in their bank account entry
    - E.g., nobody changed the content of my folder without me noticing
- Availability
  - Data or system will be available when needed
    - E.g., my website/service is protected from a malicious hacker trying to take them with a denial of service attack
- **Authenticity**
  - Belief in the source of the data
    - E.g., this new COVID study came from a reputable journal
    - E.g., the person communicating with me is who they say they are

Confidentiality: who is your adversary?

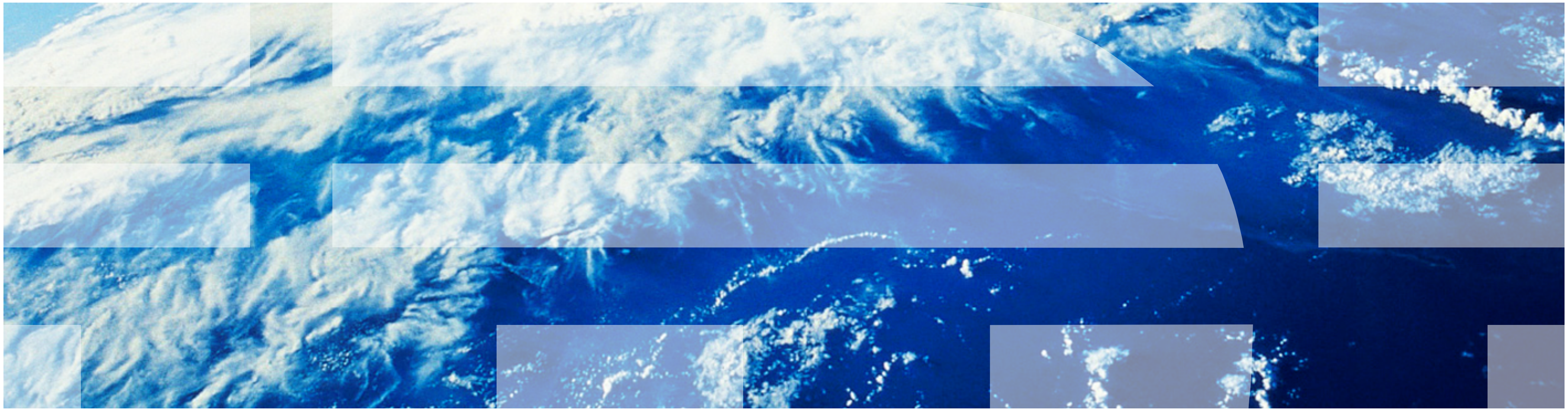
- Who are you afraid will access your data?
  - Hackers
  - Your husband/wife/boyfriend/girlfriend/mother/father…
  - Competitors
  - Anyone online
  - Your cloud provider (Google, Amazon, FB, etc.)
  - Your Internet provider
  - Your government
  - A foreign government
  - …
- → Need to explicitly define adversary

Goal: end-to-end security

- Confidentiality can be applied at any level of the stack
  - Data (e.g., encryption of data in a database, on a cloud file system)
  - Network (e.g., HTTPS)
  - Laptop/phone (e.g., full-disk encryption)
  - User identity (e.g., requiring password, multi-factor authentication)
  - …
- But applying security just at a single level does not mean your data is secure
- For example:
  - Uploading my password to a phishing site over HTTPS
  - Thief steals my phone and phone PIN, can access my phone’s disk-encrypted data

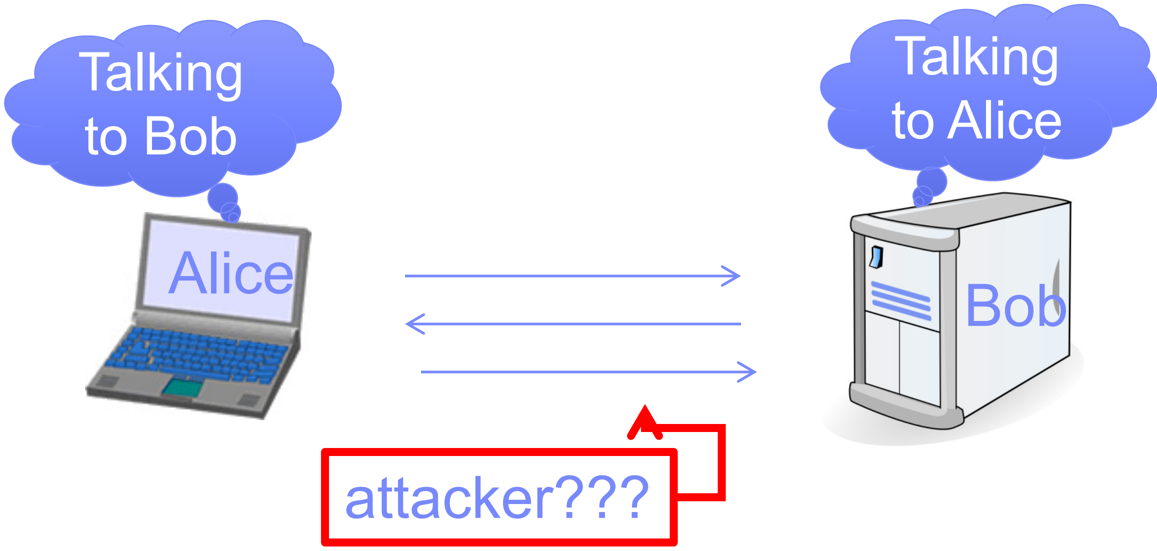
Some slides borrowed from Dan Boneh

Cryptography



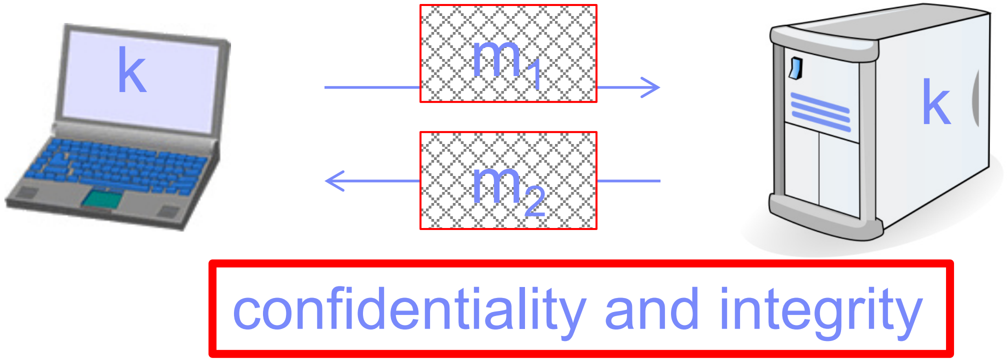
Crypto core

Secret key establishment:



attacker???

Secure communication:

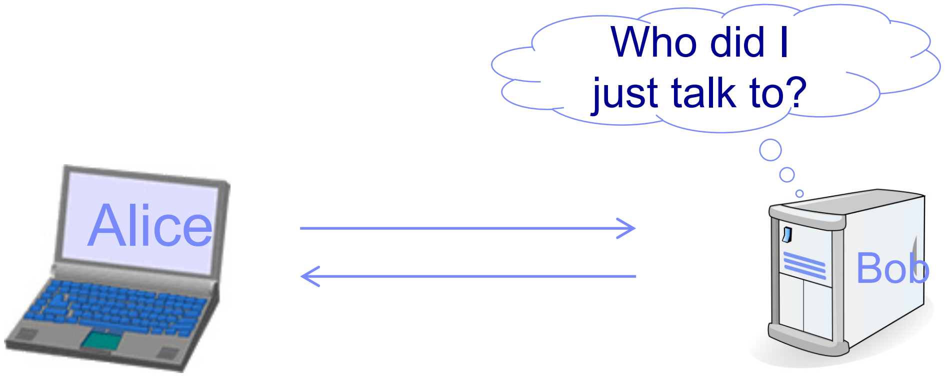


confidentiality and integrity

But crypto can do much more

- Digital signatures (authentication)

- Anonymous communication



- Cryptocurrencies and blockchain

Symmetric encryption

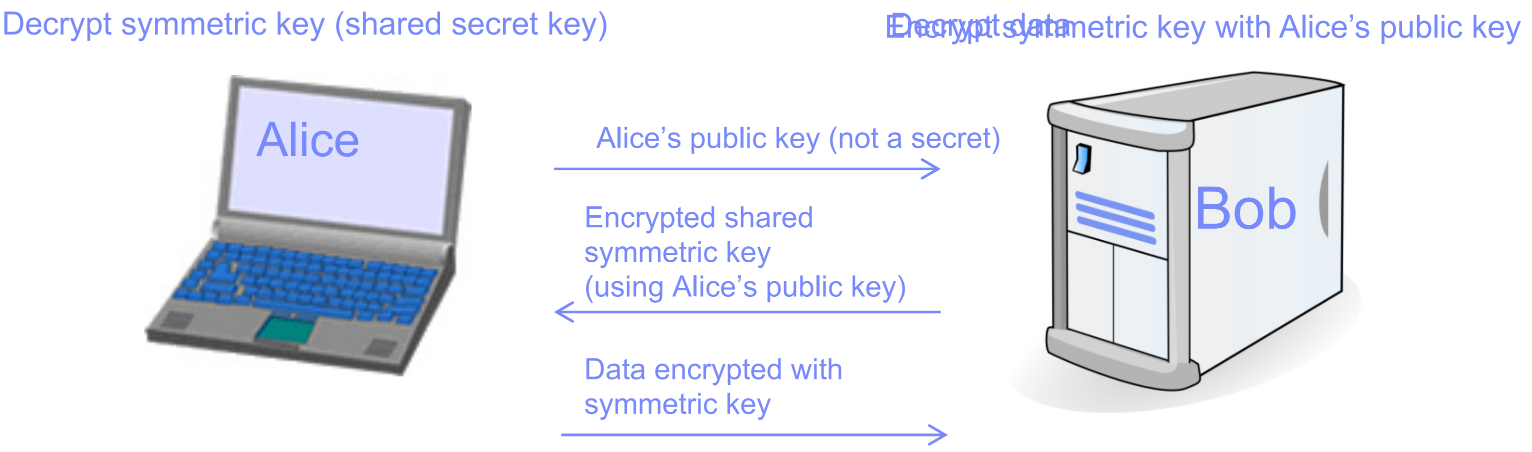
- Encryption function E(x,k)=y produces an output that appears to be totally random
  - x is unencrypted data (plaintext), k is key
- Decryption function D(y,k)=x
  - Decryption function is usually much harder to compute than encryption function
- The key is secret
  - If you obtain the key, you can decrypt the data
- Recall that hashing function h(x) produces a random output
  - Hashing is like a one-way encryption (can't be decrypted)
  - Hashing outputs can be the same for different inputs if the output space of the function is small
    - In cryptography we try to avoid that

Symmetric encryption in the real world

- Symmetric encryption is the workhorse of encryption techniques
  - Used to encrypt/decrypt bulk data (storage, network packets, etc.)
- Most common algorithm: AES
  - Use a single secret key to encrypt and decrypt
  - Usually relatively good performance
  - Modern CPUs have support for high AES throughput
- Most modern algorithms are heuristic based
  - There is no formal proof that AES decryption is hard
  - But in practice it has withstood many attempts to hack over the years

Asymmetric encryption

- Used for confidentiality, integrity and authenticity
- Similar to symmetric encryption, except encryption and decryption use different keys
- Encryption algorithm uses **public key**, which is published (not a secret)
- Decryption algorithm uses **private key**, which is a secret
- Why is it useful to use different keys for encryption and decryption?
  - For example, if two parties want to share a secret (e.g., a symmetric key) over untrusted network

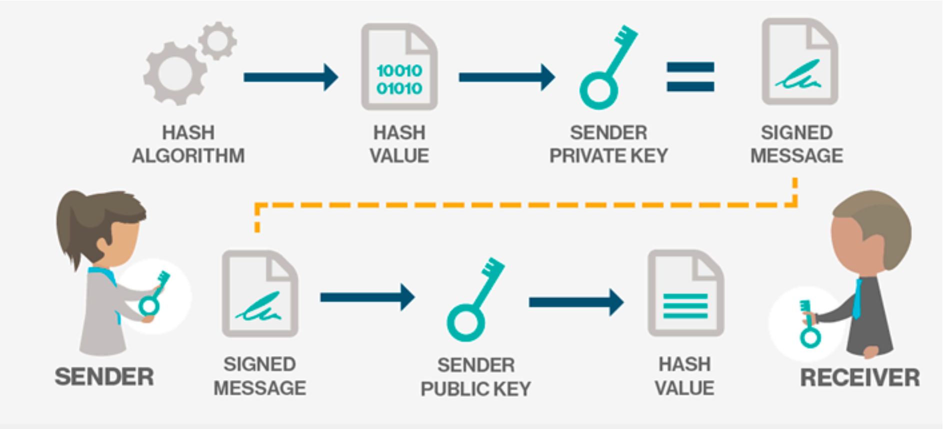

Decrypt symmetric key (shared secret key)
Encrypt symmetric key with Alice's public key
Decrypt data
Alice
Bob
Alice's public key (not a secret)
Encrypted shared symmetric key (using Alice's public key)
Data encrypted with symmetric key

Asymmetric encryption in the real world

- Asymmetric encryption is mostly used to encrypt symmetric keys or small shared secrets
  - Why? Because it usually has lower performance than symmetric encryption
- Most common algorithms: RSA
- RSA is more theoretically sound than AES
  - Based on simple yet ingenious idea that multiplication is much easier than factorization
- Basic idea:
  - Private key is based on two very large prime numbers
  - The public key is their product
- Even so, factorization problem has not been proven as “hard”

Digital signatures

- Used for integrity and authenticity (not confidentiality)
- Goal: anyone can verify the person sending the message has the private key (the secret)
- Step 1: hash your data
  - Also called a digest
- Step 2: use private key to sign message
- Append signature to your message
- Receiver can then verify using the sender's public key



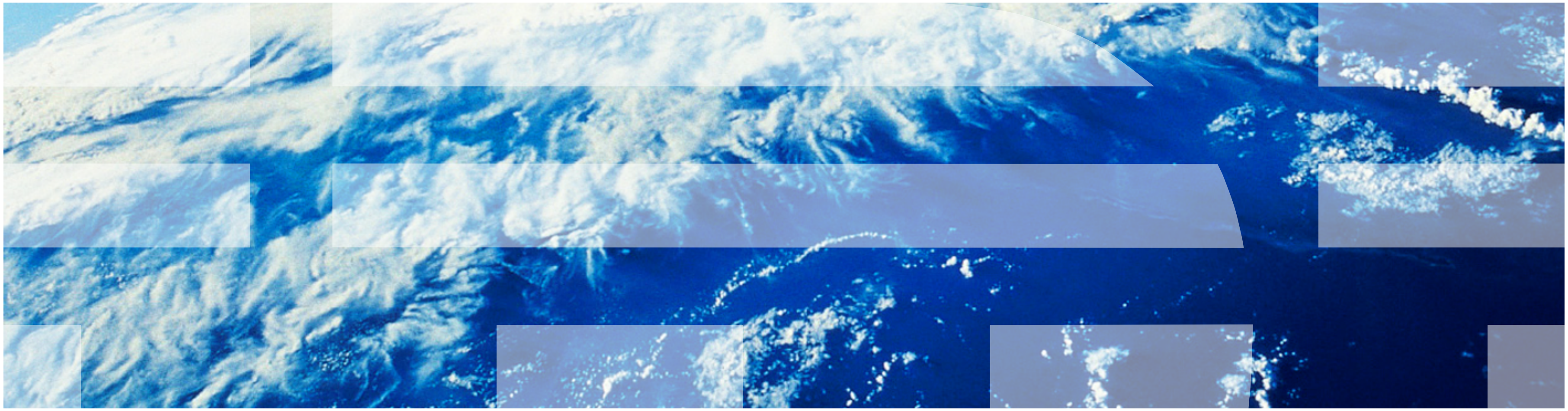
Digital signatures in the real world

- Used widely for authentication
  - This website actually belongs to google.com
  - This email was sent from columbia.edu
  - This credit card is real
- Common algorithms: DSA, RSA signatures
  - Based on asymmetric encryption
  - Generally slow

Encryption in the real world

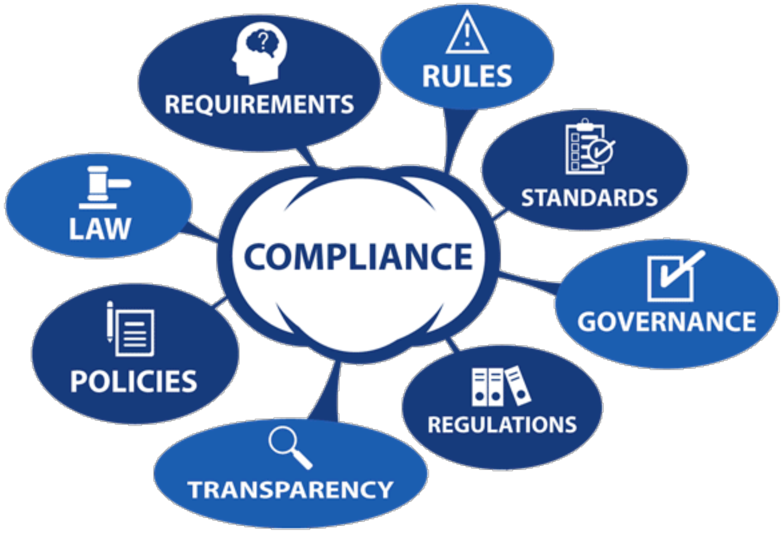
- You should never invent your own encryption algorithm or protocols
- Always use existing libraries / protocols
  - Even when using existing encryption libraries, it is tricky not to use them incorrectly
- Even widely used open source protocols have had security flaws
  - E.g., a recent example: OpenSSL Heartbleed vulnerability

Data Compliance and Privacy



Security and compliance are not the same

- Data compliance definition:
  - The process of ensuring that a dataset confirms to the rules specified by national or international laws, or the standards set by trade bodies
- Sometimes compliance includes security measures, but not always
- Compliance is primarily a legal framework on how to handle data
  - Usually more concerned with privacy than security
- Being compliant does not mean your data is secure!



Types of Data Compliance

- HIPAA
  - Ensuring the privacy of healthcare patient data
- GDPR
  - Ensuring privacy of citizens of the European Union
- PCI
  - Protecting credit card data
- FERPA
  - Protecting privacy of student information (personal details, grades, etc.)
- …

Does compliance affect you?

- Compliance affects the following use cases:
  - Dealing with personally identifiable information (PII), i.e., personal details of humans (names, locations, email addresses, phone numbers, social security numbers, etc.)
  - Healthcare
  - Finance
  - Education
  - Defense
- Usually relevant if you work on some kind of private data

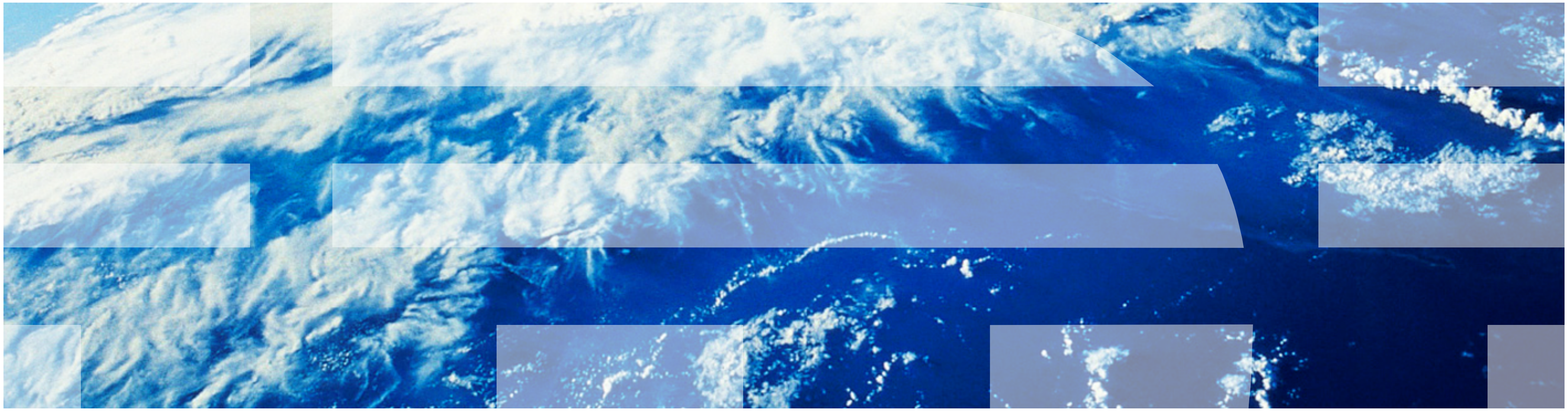
General principles

- “Protect” PII
  - This usually means some kind of encryption, but it’s not always clear at what level
- Auditability
  - Need a permanent log of any operation on the PII
    - Who read/modified/deleted, when and what did they do
  - Cannot delete log
- Access control
  - Restrictions on who can access PII
- Restrictions on sharing PII
  - Often needs to be shared securely
  - Might require the other side to sign a legal agreement
- “Right to be forgotten”
  - If user asks to delete their data, you need to do so
- Restrictions on where data can be stored
  - E.g., EU PII can only be stored in EU

How do big data systems implement compliance?

- Certain systems might be designated for storing PII, separate from other non-sensitive systems
- Require access control (who can access/update system, be able to revoke access)
- Require audit logging
- Encryption of sensitive data within the system (**encryption at rest**) and in transit when ingesting/leaving to and from the system (**encryption in transit**)
- Often separate data from different regions
  - E.g., EU storage system sits in European datacenter
- The major public cloud providers have built-in compliance controls

Course Summary



Some common themes across the class

- Covered data science / big data systems from a single node to a distributed cluster
- Different layers of the stack
  - Storage
  - File system
  - Key-value store
  - Database
- Tension between flexible API (e.g., SQL) and guarantees (e.g., ACID) and scalability
- All systems need to deal with failures
  - Things get more complicated with distributed systems
- Some common mechanisms that exist in all layers of the stack
  - Indexing
  - Filters
  - Caches
  - Replication
- Security and compliance always need to be taken into account

Some takeaways

- Focus on optimizing the frequently accessed part of your system
  - Amdahl's law
- When analyzing system performance, it's important to understand the difference between latency and throughput
  - And the importance of 99th percentile latency and stragglers for distributed systems
- Don't jump to the most complicated solution/system
  - If a single node is good enough, great!
  - If your query runs fine on a SQL server, good for you!
  - If a random forest is good enough, great!

Thank you’s

- I am extremely thankful to our awesome TAs:
  - **Hongyi Wang** (head TA) and recipient of Andrew P. Kosoresow Memorial Award for Excellence in Teaching and Service for 2020
  - **Yu Jian Wu**
  - **Qianrui Zhang**
  - **Mingen Pan**
  - **Junlin Song**
  - **Ke Li**
- Not only did they answer your questions in office hours and Piazza, and grade your exams, but they also helped write the exams and set up much of the homework assignments!
- Thanks to all of you for being engaged before the COVID-19 and especially after
  - It was a pleasure interacting with you!
- Please stay safe, and let me know if there’s any way I can help during these challenging times!